

Log-Euclidean Bag of Words for Human Action Recognition

Masoud Faraki ^{*,1}, Maziar Palhang ¹, Conrad Sanderson ^{2,3}

¹ Artificial Intelligence Laboratory, Electrical and Computer Engineering Department,
Isfahan University of Technology, Iran

² Queensland University of Technology, Brisbane, QLD 4000, Australia

³ NICTA, GPO Box 2434, Brisbane, QLD 4001, Australia

Abstract

Representing videos by densely extracted local space-time features has recently become a popular approach for analysing actions. In this paper, we tackle the problem of categorising human actions by devising Bag of Words (BoW) models based on covariance matrices of spatio-temporal features, with the features formed from histograms of optical flow. Since covariance matrices form a special type of Riemannian manifold, the space of Symmetric Positive Definite (SPD) matrices, non-Euclidean geometry should be taken into account while discriminating between covariance matrices. To this end, we propose to embed SPD manifolds to Euclidean spaces via a diffeomorphism and extend the BoW approach to its Riemannian version. The proposed BoW approach takes into account the manifold geometry of SPD matrices during the generation of the codebook and histograms. Experiments on challenging human action datasets show that the proposed method obtains notable improvements in discrimination accuracy, in comparison to several state-of-the-art methods.

Keywords: Riemannian manifolds, Region covariance descriptor, Bag of Words, Human action recognition.

1 Introduction

Among several video analysis tasks, human action recognition has received significant attention, mainly because of its applications to visual surveillance, content-based video analysis, and human-computer interaction [1, 31, 34, 46, 41]. Many methods have been proposed for reliable action recognition based on various feature detectors/descriptors to capture local motion patterns [9, 18, 16, 38, 48, 45, 44]. Dense space-time representation of videos has been recently shown to be promising for the action categorisation task [45, 44]. This in turn suggests the need for employing descriptors to compactly represent the dense collection of local features.

In this paper, we utilise region covariance matrices, composed from densely sampled features, as the descriptors. Such use of covariance matrices as image descriptors is relatively novel. They were introduced by Tuzel et al. [42] and since then have been employed successfully for pedestrian detection [43], non-rigid object tracking [32], face recognition [29], and analysing diffusion tensor images [30]. Furthermore, a spatio-temporal version of covariance matrix descriptors has shown superior performance for action/gesture recognition [36].

Utilising a covariance matrix as a region descriptor has several advantages. Firstly, it captures second-order statistics of the local features. Secondly, it is straightforward approach to fusing various (correlated) features.

Thirdly, it is a low dimensional descriptor and is independent of the size of the region. Fourth, through the averaging process in its computation, the impact of the noisy samples is reduced. Finally, efficient methods for its fast computation in images and videos are available [43, 36]. While the above advantages make covariance-based descriptors attractive, using them for discrimination purposes can be challenging. Covariance matrices are Symmetric Positive Definite (SPD) matrices and naturally form a connected Riemannian manifold. This can make inference methods based on covariance matrices more difficult, as manifold curvature needs to be taken into account [30, 43].

Within the fields of image categorisation and face recognition it has been shown that discrimination approaches based on Bag of Words (BoW) are effective [28, 35, 47, 49]. In a traditional BoW approach, a set of low-level descriptors is typically encoded as a high-dimensional histogram, with each entry in the histogram representing a count or probability of occurrence of a ‘visual’ word. The dictionary (or codebook) of words is fixed and obtained during a training stage, typically through k -means clustering. The resulting histograms are then interpreted as medium-level feature vectors and fed into standard classifiers.

Contributions. Following the trend of adapting machine learning tools originally designed for vector spaces to their Riemannian counterparts [30, 43, 40, 39, 14, 13, 36, 51], in this work we propose to extend the general BoW approach to handle covariance matrices that are treated as points on a Riemannian manifold. We first form spatio-temporal covariance descriptors from densely extracted motion-based features, namely Histograms of Optical Flow (HOF) introduced by Laptev et al. [20]. The covariance descriptors are then encoded in a Log-Euclidean Bag of Words (LE-BoW) model. To achieve this, we use a diffeomorphism and form the LE-BoW model by embedding the Riemannian manifold into a vector space. The embedding is obtained by flattening the manifold through tangent spaces. We explore several encoding methods within the LE-BoW framework. We then compare and contrast the proposed approach against recent action recognition methods proposed by Wang et al. [44], Messing et al. [26], and Niebles et al. [27]. Empirical results on three datasets (KTH [37], Olympic Sports [27], Activity of Daily Living [26]) show that the proposed action recognition approach obtains superior performance.

We continue this paper as follows. Section 2 provides an overview of recent work in action recognition. Section 3 is dedicated to Riemannian geometry and serves as a grounding for following sections. Section 4 discusses the LE-BoW model. In Section 5 we compare the performance of the proposed method with previous approaches on several datasets. The main findings and future directions are summarised in Section 6.

2 Related Work

Human action recognition has been addressed extensively in the computer vision community from various perspectives. Some methods rely on global descriptors; two examples are the methods proposed by Ali and Shah [3] and Razzaghi et al. [33]. In [3], a set of optical flow based kinematic features is extracted. Kinematic models are computed by applying principal component analysis on the volumes of kinematic features. Razzaghi et al. [33] represent human motion by spatio-temporal volume and propose a new affine invariant descriptor based on a function of spherical harmonics. A downside of global representations is their reliance on localisation of the region of interest, and hence they are sensitive to viewpoint change, noise, and occlusion [36].

To address the abovementioned issues, videos of actions can also be represented through sets of local features, either in a sparse [37, 9] or dense [45, 44] manner. Sparse feature detectors (also referred to as interest point detectors) abstract video information by maximising saliency functions at every point in order to extract salient spatio-temporal patches. Examples are Harris3D [19] and Cuboid [9] detectors. Laptev and Lindeberg [19] extract

interest points at multiple scales using a 3D Harris corner detector and subsequently process the extracted points for modelling actions. The Cuboid detector proposed by Dollar et al. [9] extracts salient points based on temporal Gabor filters. It is especially designed to extract space-time points with local periodic motions.

Wang et al. [45] demonstrate that dense sampling approaches consistently outperform space-time interest point based methods for human action categorisation. A dense sampling at regular positions in space and time guarantees good coverage of foreground motions as well as of surrounding context. To characterise local patterns (i.e. motion, appearance, or shape), the descriptors divide small 3D volumes into a grid of $n_x \times n_y \times n_t$ cells and for each cell the related information is accumulated. Examples are HOG and HOF [20], HOG3D [16], and 3D SIFT [38].

An alternative line of research proposes to track given spatial point over time and capture related information. Messing et al. [26] track Harris3D [19] interest points with a KLT tracker [23] and extract velocity history information. To improve performance, other useful features such as appearance and location are taken into account in a generative mixture model. Recently, Wang et al. [44] show promising results by tracking densely sampled points and extract aligned shape, appearance, and motion features. They also introduce Motion Boundary Histograms (MBH) based on differential optical flow.

3 Riemannian Geometry

In this section, we review Riemannian geometry on the manifold of real SPD matrices. We first formally define a covariance matrix descriptor for the whole video. Let $\mathbb{I} = \{I_t\}_{t=1}^T$ denote a set of $W \times H$ greyscale frames of a video. Also, let $\mathbb{O} = \{\mathbf{o}_i\}_{i=1}^n$ be a set of observations $\mathbf{o}_i \in \mathbb{R}^d$ extracted from \mathbb{I} . For example, one might extract a d dimensional feature vector at each pixel, resulting in $W \times H \times T$ observations. Then, \mathbb{I} can be represented by a $d \times d$ covariance matrix of the observations as:

$$\begin{aligned} C_I &= \frac{1}{n-1} \sum_{i=1}^n (\mathbf{o}_i - \boldsymbol{\mu})(\mathbf{o}_i - \boldsymbol{\mu})^T, \\ \boldsymbol{\mu} &= \frac{1}{n} \sum_{i=1}^n \mathbf{o}_i. \end{aligned} \quad (1)$$

The entries on the diagonal of matrix C_I are the variances of each feature and the non-diagonal entries are their pairwise correlations (see Fig. 1 for a conceptual diagram). There are several reasons as to why covariance matrices are attractive for representing images and videos: **(i)** they provide a natural way for fusing various features; **(ii)** they can reduce the impact of noisy samples through the averaging operation in its computation; **(iii)** a $d \times d$ covariance matrix is usually low-dimensional and independent of the size of the region; **(iv)** they can be efficiently computed using integral images/videos [43, 36]; **(v)** affine invariant metrics exist to compare covariance matrices [30].

A *manifold*, \mathcal{M} , is a locally Euclidean topological space. Locally Euclidean means that each point has some neighbourhood that is homeomorphic (one-to-one, onto, and continuous in both directions) to an open ball in \mathbb{R}^d , for some d . The *tangent space* at a point \mathbf{P} on the manifold, $T_{\mathbf{P}}\mathcal{M}$, is a vector space that consists of the tangent vectors of all possible curves passing through \mathbf{P} (see Fig. 2 for an illustration). On the manifold, a *Riemannian metric* is defined as a continuous collection of dot products on the tangent space $T_{\mathbf{P}}\mathcal{M}$ at each $\mathbf{P} \in \mathcal{M}$. The Riemannian metric of the manifold enables us to define geometric notions on the manifold such as lengths and angles. The *geodesic distance* between two points on the manifold is defined as the length of the shortest curve connecting the two points.

A *Riemannian manifold* (\mathcal{M}, g) consists of the analytic manifold \mathcal{M} and its associated metric $g_{\mathbf{P}}(\cdot, \cdot) : \mathcal{M} \times \mathcal{M} \rightarrow \mathcal{R}$ that varies smoothly on $T_{\mathbf{P}}\mathcal{M}$. The function g has a symmetric, positive definite bi-linear form on each

$\mathbf{p} \in T_P \mathcal{M}$. It can be chosen to provide robustness to some geometrical transformations.

Two operators, namely the *exponential map* $\exp_P(\cdot) : T_P \mathcal{M} \rightarrow \mathcal{M}$ and the *logarithm map* $\log_P(\cdot) = \exp_P^{-1}(\cdot) : \mathcal{M} \rightarrow T_P \mathcal{M}$, are defined over differentiable manifolds to switch between the manifold and tangent space at P . The exponential operator maps a tangent vector Δ to a point X on the manifold. The property of the exponential map ensures that the length of Δ becomes equal to the geodesic distance between X and P . The logarithm map is the inverse of the exponential map and maps a point on the manifold to the tangent space $T_P \mathcal{M}$. The exponential and logarithm maps vary as point P moves along the manifold. We refer interested readers to [5, 24] for more detailed treatment on manifolds and related topics.

3.1 Riemannian Manifold of SPD Matrices

The space of real $d \times d$ SPD matrices, \mathcal{S}_{++}^d , forms a Lie Group which is an algebraic group with a manifold structure. It is natural to use the language of Riemannian manifolds and all the related concepts of differential geometry when discussing \mathcal{S}_{++}^d .

The Affine Invariant Riemannian Metric (AIRM) [30] on \mathcal{S}_{++}^d is defined as:

$$\langle \mathbf{v}, \mathbf{w} \rangle_P := \langle \mathbf{P}^{-1/2} \mathbf{v} \mathbf{P}^{-1/2}, \mathbf{P}^{-1/2} \mathbf{w} \mathbf{P}^{-1/2} \rangle = \text{tr}(\mathbf{P}^{-1} \mathbf{v} \mathbf{P}^{-1} \mathbf{w}), \quad (2)$$

for $\mathbf{P} \in \mathcal{S}_{++}^d$ and $\mathbf{v}, \mathbf{w} \in T_P \mathcal{M}$, induces the following geodesic distance between points $\mathbf{X}, \mathbf{Y} \in \mathcal{S}_{++}^d$:

$$\delta_R(\mathbf{X}, \mathbf{Y}) = \|\log(\mathbf{X}^{-1/2} \mathbf{Y} \mathbf{X}^{-1/2})\|_F. \quad (3)$$

For the AIRM, the logarithm and exponential maps are given by [5]:

$$\log_P(\mathbf{X}) = \mathbf{P}^{\frac{1}{2}} \log(\mathbf{P}^{-\frac{1}{2}} \mathbf{X} \mathbf{P}^{-\frac{1}{2}}) \mathbf{P}^{\frac{1}{2}}, \quad (4)$$

$$\exp_P(\mathbf{X}) = \mathbf{P}^{\frac{1}{2}} \exp(\mathbf{P}^{-\frac{1}{2}} \mathbf{X} \mathbf{P}^{-\frac{1}{2}}) \mathbf{P}^{\frac{1}{2}}. \quad (5)$$

In Eqns. (4) and (5), $\log(\cdot)$ and $\exp(\cdot)$ are the matrix logarithm and exponential operators, respectively. For SPD matrices, they can be computed through Singular Value Decomposition (SVD). If we let $\text{diag}(\lambda_1, \lambda_2, \dots, \lambda_d)$ be a diagonal matrix formed from real values $\lambda_1, \lambda_2, \dots, \lambda_d$ on diagonal elements and $\mathbf{X} = \mathbf{U} \text{diag}(\lambda_i) \mathbf{U}^T$ be

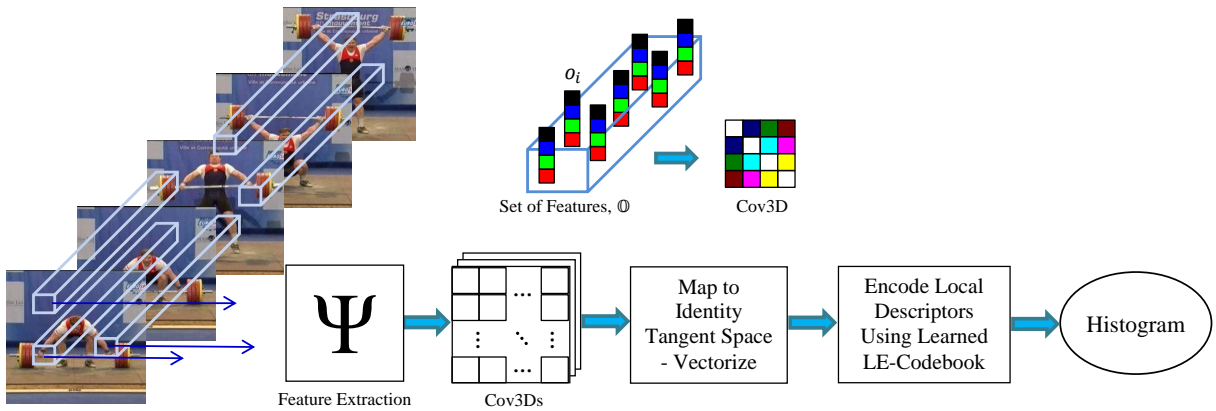


Figure 1: Conceptual block diagram showing computations of LE-BoW histogram generation.

the SVD of the symmetric matrix \mathbf{X} , then

$$\log(\mathbf{X}) = \sum_{r=1}^{\infty} \frac{(-1)^{r-1}}{r} (\mathbf{X} - \mathbf{I})^r = \mathbf{U} \text{diag}(\ln(\lambda_i)) \mathbf{U}^T, \quad (6)$$

$$\exp(\mathbf{X}) = \sum_{r=0}^{\infty} \frac{1}{r!} \mathbf{X}^r = \mathbf{U} \text{diag}(\exp(\lambda_i)) \mathbf{U}^T. \quad (7)$$

4 Log-Euclidean Bag of Words

In this section we discuss how a conventional Bag of Words (BoW) model can be extended to incorporate the Riemannian structure of covariance matrices. In a nutshell, the BoW representation is obtained by first clustering a large set of selected local descriptors with (usually) k -means, in order to acquire a visual vocabulary or codebook. Then, a histogram is extracted by assigning each descriptor to its closest visual word.

To devise a BoW model on Riemannian manifolds, we should address two sub-problems:

1. Given a set of training samples $\mathbb{X} = \{\mathbf{X}_i\}_{i=1}^N$ from the underlying \mathcal{S}_{++}^d manifold (where each point on the manifold corresponds to a covariance matrix), how can a codebook $\mathbb{D} = \{\mathbf{D}_j\}_{j=1}^k$ be obtained?
2. Given a codebook $\mathbb{D} = \{\mathbf{D}_j\}_{j=1}^k$ and a set of covariance matrices $\mathbb{Q} = \{\mathbf{Q}_i\}_{i=1}^p$ extracted from a query video, how can a histogram be obtained for classification?

4.1 Riemannian Codebook

In the most straightforward case, one can neglect the geometry of SPD matrices and vectorise training data to learn a codebook. We note that SPD matrices form a closed set under normal matrix addition, i.e., adding two SPD matrices results in another SPD matrix. Therefore, a codebook can be generated by applying k -means on vectorised data. More specifically, the resulting clusters are determined by computing the arithmetic mean of the nearest training vectors to that cluster.

Despite its simplicity, several studies argue against exploiting Euclidean geometry and vector form of SPD matrices for inference [30, 43]. For instance, as shown by Pennec [30] the determinant of the weighted mean could become greater than samples' determinants, an undesirable outcome known as the swelling effect [4]. Moreover, symmetric matrices with negative or zero eigenvalues are at a finite distance from any SPD matrix in this framework. In many problems like diffusion tensor MRI, this is not physically acceptable [4, 30]. Therefore, geometry of SPD matrices should be considered in creating the codebook.

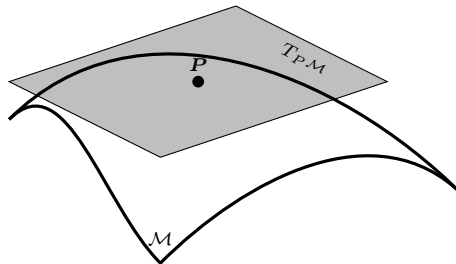


Figure 2: Conceptual illustration of the tangent space at point P on a Riemannian manifold \mathcal{M} .

To benefit from Riemannian geometry, an alternative is to replace the arithmetic mean with Karcher mean (also referred as Fr chet or Riemannian mean) [30]. The Karcher mean is the point that minimises the following metric dispersion:

$$\mathbf{X}^* = \arg \min_{\mathbf{X}} \sum_{i=1}^N \delta_g^2(\mathbf{X}_i, \mathbf{X}), \quad (8)$$

where $\delta_g : \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R}^+$ is the associated geodesic distance function. The discussion of the existence and uniqueness value of the Karcher mean as well as its computation are given in [30].

Computing the Karcher mean requires switching back and forth between a manifold and its tangent spaces. This is computationally demanding, especially in our application where a large number of high dimensional training points is available. More precisely, each mapping to a tangent space can be computed using Cholesky factorisation with $O(d^3)$ for a $d \times d$ covariance matrix. Therefore, we opt for a faster way of computing a codebook by minimum use of the logarithm map, i.e., Eqn. (4).

Our idea here is to simplify the problem by embedding the manifold into a vector space. For this purpose, we make use of a mapping from \mathcal{S}_{++}^d into the space of symmetric matrices by the principal matrix logarithm. The motivation comes from the fact that unlike the general case of invertible square matrices, there always exists a unique, real and symmetric logarithm for any SPD matrix, which can be obtained by principal logarithm. Moreover, $\log(\cdot)$ on \mathcal{S}_{++}^d is diffeomorphism (a one-to-one, continuous, differentiable mapping with a continuous, differentiable inverse). Formally,

Theorem 1. $\log(\cdot) : \mathcal{S}_{++}^d \rightarrow \text{Sym}(d)$ is C^∞ and therefore both $\log(\cdot)$ and its inverse $\exp(\cdot)$ are smooth, i.e., they are diffeomorphisms.

Proof. We refer the reader to [4] for the proof of this theorem. □

Embedding into the space of $d \times d$ symmetric matrices, $\text{Sym}(d)$, through principal logarithm can be also understood as embedding \mathcal{S}_{++}^d into its tangent space at identity matrix. Since symmetric matrices (or equivalently tangent spaces) form a vector space, then we can seamlessly employ Euclidean tools (like k -means to obtain a codebook) to tackle the problem in hand. Other properties of the induced space, the log-Euclidean space, are studied in [4]. We note that our idea here can be labelled as an extrinsic approach, i.e., it depends on the embedding Euclidean space.

Given an SPD matrix \mathbf{X} , its log-Euclidean vector representation, $\mathbf{a} \in \mathbb{R}^m, m = \frac{d(d+1)}{2}$, is unique and defined as $\mathbf{a} = \text{Vec}(\log(\mathbf{X}))$ where $\text{Vec}(\mathbf{B}), \mathbf{B} \in \text{Sym}(d)$ is:

$$\text{Vec}(\mathbf{B}) = [b_{1,1}, \sqrt{2}b_{1,2}, \sqrt{2}b_{1,3}, \dots, \sqrt{2}b_{1,d}, b_{2,2}, \sqrt{2}b_{2,3}, \dots, b_{d,d}]^T. \quad (9)$$

Having the training samples mapped to the identity tangent space, we seek to estimate k clusters C_1, C_2, \dots, C_k with centers $\{\mathbf{D}_j\}_{j=1}^k$ such that the sum of distances over all clusters is minimised. This can be solved using the conventional k -means algorithm [6]. The procedure is summarised in Algorithm 1.

4.2 Encoding Local Descriptors

In the previous section, we elaborated on how a codebook for covariance matrices can be obtained. In this subsection, we elaborate on several encoding methods for a set of local descriptors. In other words, having a codebook, $\mathbb{D} = \{\mathbf{D}_j\}_{j=1}^k$, at our disposal (obtained by Algorithm 1), we seek to group a set of covariance matrices,

Algorithm 1 Log-Euclidean k-means algorithm over \mathcal{S}_{++}^d for learning the visual dictionary

Input:

- training set $\mathbb{X} = \{\mathbf{X}_i\}_{i=1}^N$ from the underlying \mathcal{S}_{++}^d manifold
- $nIter$, the number of iterations

Output:

- Visual dictionary $\mathbb{D} = \{\mathbf{D}_j\}_{j=1}^k$, $\mathbf{D}_j \in \mathbb{R}^m$

```

1: Compute  $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^N$ , log-Euclidean representation of  $\mathbb{X}$  using  $\mathbf{x}_i = \text{Vec}(\log(\mathbf{X}_i))$ .
2: Initialise the dictionary  $\mathbb{D} = \{\mathbf{D}_j\}_{j=1}^k$  by selecting  $k$  samples from  $\mathbf{x}$  randomly.
3: for  $t = 1 \rightarrow nIter$  do
4:   Assign each point  $\mathbf{x}_i$  to its nearest cluster in  $\mathbb{D}$ .
5:   Compute the average dispersion from cluster centers by  $\varepsilon = \frac{1}{N} \sum_{j=1}^k \sum_{\mathbf{x}_i \in C_j} \text{dist}(\mathbf{x}_i, \mathbf{D}_j)$ .
6:   If  $\varepsilon$  is less than a predefined threshold, then break the loop; else recompute cluster centres  $\{\mathbf{D}_j\}_{j=1}^k$  by  $\mathbf{D}_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i$ .
7: end for
  
```

$\mathbb{Q} = \{\mathbf{Q}_i\}_{i=1}^p$, extracted from a query video, in order to find a histogram based representation. Similar to the code-book learning stage, we first compute the log-Euclidean representation of \mathbb{Q} using $\mathbf{q}_i = \text{Vec}(\log(\mathbf{Q}_i))$, $\mathbf{q}_i \in \mathbb{R}^m$. Fig. 1 shows a conceptual diagram of our proposed histogram generation approach.

There are several ways of obtaining a histogram based representation, ranging in terms of complexity and amount of spatial and/or temporal information retained. In this work we evaluate three methods, elucidated in the following subsections: **(i)** hard assignment, **(ii)** spatio-temporal pyramids, **(iii)** sparse coding.

4.2.1 Hard Assignment (HA)

In its most straightforward and simplest form, for the set $\{\mathbf{q}\}$, a histogram H is obtained by Hard Assignment (HA), which is related to Vector Quantisation [49]. This requires $p \times k$ comparisons. The j -th ($1 \leq j \leq k$) dimension of H is obtained using $H_j = |C_j|$, where $|C_j|$ denotes the number of vectors \mathbf{q}_i ($1 \leq i \leq p$) assigned to the j -th cluster. The vectors are assigned to their closest vocabulary word in the dictionary using Euclidean distance. The resulting histogram is ℓ_2 normalised via $\hat{H} = \frac{H}{\|H\|_2}$.

4.2.2 Spatio-Temporal Pyramids (STP)

HA encoding loses structure information between the vectors in the set $\{\mathbf{q}\}$. We encode the structure information to our LE-BoW model by incorporating Spatio-Temporal Pyramids (STP) [44, 20], an extension of spatial pyramids [21, 47]. For spatial domain we use the entire spatial block, a subdivision into three horizontal stripes, and a 2×2 spatial grid. For the temporal domain we use the entire duration as well as a subdivision into 2 temporal blocks. For each cell of the grid, a separate hard assigned LE-BoW histogram is computed. Then, a video is represented as concatenation of the cell histograms. We use each grid structure as a separate channel and combine them using a χ^2 kernel (see Section 5). As illustrated in Fig. 3, we have six channels to represent a video in STP encoding.

For classification, we use a non-linear support vector machine with a multi-channel RBF- χ^2 kernel defined by:

$$K(H_i, H_j) = \exp \left(- \sum_c \frac{1}{A^c} \delta_{\chi^2}(H_i^c, H_j^c) \right), \quad (10)$$

where $\delta_{\chi^2}(H_i^c, H_j^c)$ is the χ^2 distance between histogram H_i and H_j with respect to the c -th channel, and A^c is the mean value of the χ^2 distances between the training samples for the c -th channel.

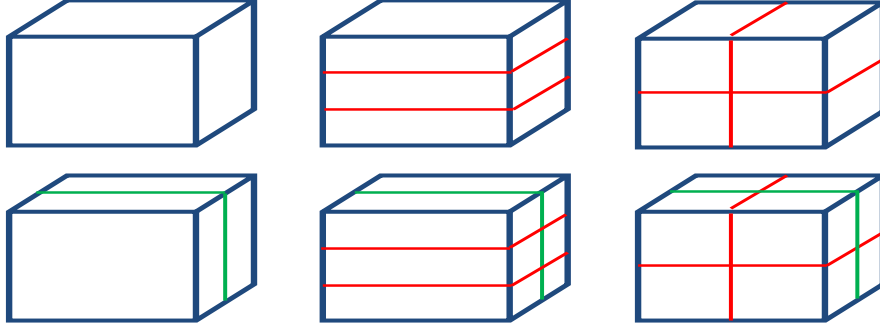


Figure 3: Spatio-temporal grids for STP histogram encoding.

4.2.3 Sparse Coding (SC)

Sparse Coding (SC), the optimal linear decomposition of a signal using a few elements of a dictionary has proved to be effective for various computer vision tasks [10, 50, 49]. Since the resulting histogram by either HA or STP is naturally sparse, it is possible to employ SC algorithms to encode local descriptors. We use Algorithm 1 to train a dictionary for SC. However, it is also possible to use dedicated algorithms for this purpose [2, 17].

Kernel sparse coding was previously proposed in [13] to take into account the geometry of SPD matrices with the aid of the Stein kernel [8]. However, the Stein metric fails in our application where many low rank SPD matrices exist. More specifically, the determinant of SPD matrices formed from HOF features can be close to zero. As a result, other SPD matrices locate at infinite distance to those low rank matrices.

A vector of weights $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_k]^T$ is computed for each $q_i \in \mathbb{R}^m, 1 \leq i \leq p$, by solving a minimisation problem that selects a sparse set of dictionary atoms. More specifically, having a dictionary $D \in \mathbb{R}^{m \times k}$ at our disposal, the weight vector $\alpha \in \mathbb{R}^k$ is obtained via solving the following ℓ_1 -minimisation (also known as Lasso [10]):

$$\min \frac{1}{2} \|D\alpha - q_i\|_2^2 + \lambda \|\alpha\|_1. \quad (11)$$

Pooling local sparse codes is performed via averaging. To solve Eqn. (11) we used the SPAMS optimisation toolbox (<http://spams-devel.gforge.inria.fr>) for sparse estimation problems.

4.3 Computational Complexity

The covariance matrices can be computed efficiently (i.e., in one pass over the video) via integral videos [36]. This results in $O(WHTd^2)$ operations for computing a $d \times d$ covariance matrix from a $W \times H \times T$ video. Generating histograms in LE-BoW method requires covariance matrices to be mapped to log-Euclidean space first. The matrix logarithm can be computed using Cholesky factorisation with $O(d^3)$ operations. Computing K distances using Euclidean distance can be done at the cost of $O(\frac{1}{2}d^2)$. Therefore, computing a K dimensional LE-BoW (with HA encoding) signature for one covariance matrix requires $O((WHT + \frac{K}{2})d^2 + d^3)$ operations.

5 Experiments

In this section we compare and contrast the performance of the proposed LE-BoW method against several state-of-the-art approaches. Before delving into experiments, we elaborate how a descriptive representation of action videos can be attained by covariance matrices. To this end, from each video a set of covariance matrices is extracted and then passed to LE-BoW to generate histograms (see Fig. 1).

To generate covariance matrices, a set of overlapping spatio-temporal blocks are extracted from the image sequence and the covariance matrix for each block is obtained from Histogram of Optical Flow (HOF) features of densely extracted trajectories within that block. To obtain trajectories, images of a sequence are first resized to 240×360 and then pixels of an image sequence are sampled on a $W \times W$ spaced grid. Then, the location of the sampled points is estimated/tracked in L subsequent frames using the estimated optical flow field of the sequence, ω , convolved with a 3×3 median filter kernel M . More specifically, given a sampled point P_t in frame I_t , its tracked point P_{t+1} in frame I_{t+1} is obtained via $P_t + M * \omega$. Once the trajectory points $(P_t, P_{t+1}, \dots, P_{t+L-1})$ in L subsequent frames is found, the HOF is computed in an $N \times N$ pixels volume around each P_t . To embed structure information, each volume is further divided into a spatio-temporal grid of size $n_\sigma \times n_\sigma \times n_\tau$.

We have used the code available by Wang et al. [44] for our dense trajectory feature extraction and followed the default parameter values there (i.e. $W = 5$, $L = 15$, and $N = 32$). Trajectories are extracted in 8 spatial scales with $n_\sigma = n_\tau = 2$. Since each HOF is 72 dimensional (i.e. 9 bins in $n_\sigma \times n_\sigma \times n_\tau$ grid), our covariance matrices are 72×72 dimensional. To avoid having rank deficient covariance matrices, blocks with the number of trajectories below 72 are rejected. We cluster a subset of 30K randomly selected covariance matrices and fix the number of visual words to 2000. For classification, we use one-against-all approach and a non-linear support vector machine with a RBF- χ^2 kernel. We report our LE-BoW model with Hard Assignment (HA), Spatio-Temporal Pyramids (STP), and Sparse Coding (SC) encoding methods. We show the discrimination power of our proposed method against several state-of-the-art methods previously applied on three datasets: KTH [37], Olympic Sports [27], and Activity of Daily Living [26].

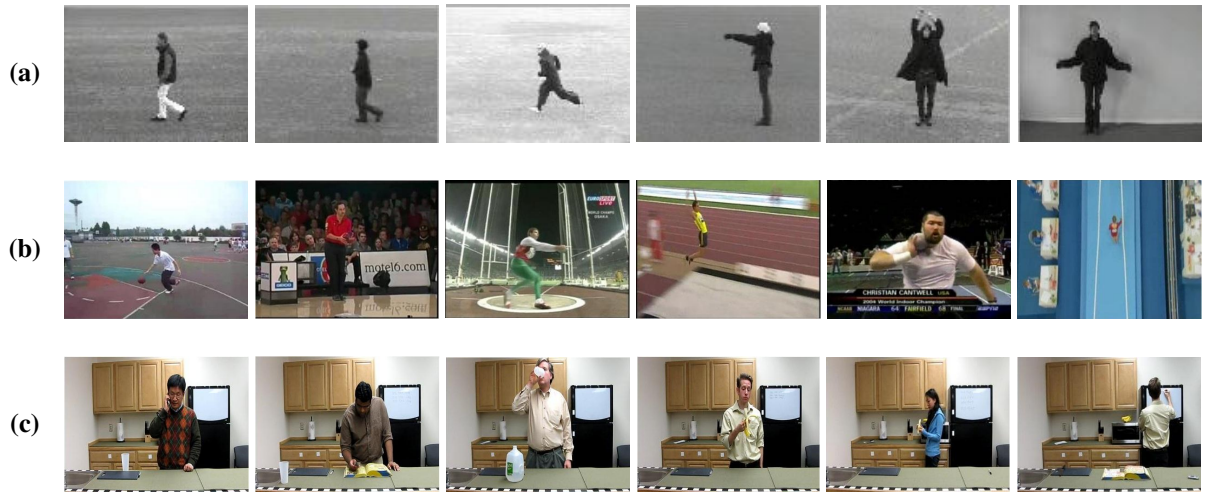


Figure 4: Example images from the datasets used in our experiments: (a) KTH [37], (b) Olympic Sports [27], (c) Activity of Daily Living [26].

Table 1: Comparison between the proposed approach against previous methods on the KTH dataset; CCR: Correct Classification Rate (in %).

| Method | CCR |
|------------------------------|-------------|
| Laptev et al. [20] | 91.8 |
| Gilbert et al. [11] | 94.5 |
| Wang et al. [44] | 95.3 |
| proposed LE-BoW (HA) | 95.0 |
| proposed LE-BoW (STP) | 95.7 |
| proposed LE-BoW (SC) | 97.4 |

5.1 KTH Dataset

The KTH dataset [37] contains six human action classes: walking, jogging, running, boxing, hand-waving, and hand-clapping, performed by 25 subjects in 4 scenarios: outdoors, outdoors with scale variation, outdoors with varying clothes, and indoors; see Fig. 4 for examples. The videos are recorded with static and homogeneous background. However, the camera is not static, i.e. vibration and unknown zooming exist. In total, the data consists of 2391 video samples. We follow the original experiment setup of the authors (i.e., dividing the samples into subjects: 2, 3, 5, 6, 7, 8, 9, 10, 22 for the test set and the remaining 16 subjects for the training set).

On KTH dataset, Laptev et al. [20] proposed a system where spatio-temporal interest points are extracted and described using HOG/HOF descriptors. In order to classify a query video, BoW model is utilised in a multi-channel SVM classifier with χ^2 kernel. Gilbert et al. [11] propose to use an overcomplete set of simple 2D corners in space and time. The extracted points are first grouped spatially and temporally using a hierarchical process. The most distinctive and descriptive features are learned. Wang et al. [44] track densely sampled points by a median filter kernel and extract aligned shape, appearance, and motion features. BoW model is utilised in a 30-channel (6 channels and 5 types of features) SVM classifier with χ^2 kernel for classification.

In Table 1, we compare our proposed method against the aforementioned methods on the KTH dataset. Our LE-BoW approach is superior to the method proposed by Laptev et al. and Gilbert et al.. The HA and STP encoding methods result in performance on par with Wang’s system. However, the SC encoding method improved the accuracy by approximately 2 percentage points.

5.2 Olympic Sports Dataset

The Olympic Sports dataset [27] contains videos of athletes practising various sport activities. All video sequences were collected from YouTube and their class labels annotated with the help of Amazon Mechanical Turk. There are 16 sports actions: high-jump, long-jump, triple-jump, pole-vault, basketball lay-up, bowling, tennis-serve, platform, discus, hammer, javelin, shot-put, springboard, snatch, clean-jerk, and vault, represented by a total of 783 videos. We use the standard train/test split recommended by the authors (649 sequences for training and 134 sequences for testing). Example images are shown in Fig. 4.

Niebles et al. [27] represent activities as temporal compositions of motion segments. They train a discriminative model that encodes a temporal decomposition of video sequences and appearance models for each motion segment. For classification, a query video is matched to the model according to the learned appearances and motion segment decomposition. The classification is based on the quality of matching between the motion segment classifiers and the temporal segments in the query sequence. In Liu et al. [22], human actions are represented by a set of action

Table 2: Comparisons between the proposed approach to the state-of-the-art methods on Olympic Sports dataset; MAP: Mean Average Precision over all classes (in %).

| Method | MAP |
|------------------------------|-------------|
| Niebles et al. [27] | 72.1 |
| Liu et al. [22] | 74.4 |
| Wang et al. [44] | 77.2 |
| proposed LE-BoW (HA) | 74.9 |
| proposed LE-BoW (STP) | 80.6 |
| proposed LE-BoW (SC) | 79.9 |

attributes. A unified framework introduced wherein the attributes can be discriminatively selected. The framework is built upon a latent SVM formulation where latent variables capture the degree of importance of each attribute for each action class.

In Table 2, we compare our proposed method against state-of-the-art methods on the Olympic Sports dataset. Mean Average Precision over all classes is reported as in [27]. Using spatio-temporal pyramids consistently improved the classification rate (from 74.9% to 80.6%). The same improvement is also observed in the work of Wang et al. [44] by considering trajectory shape, HOG, HOF, MBHx, and MBHy descriptors in a STP encoding approach. However, Wang et al. report 58.7% by single HOF descriptor. In contrast, we observed considerable improvement by taking the covariance of HOF features (i.e. 74.9% by HA). On this dataset, the SC encoding method works on par with the STP encoding approach.

5.3 Activity of Daily Living Dataset

This dataset consists of 150 videos of 5 subjects performing a series of daily tasks in a kitchen environment, acquired using a stationary camera [26]. As recommended by [26], we evaluate our results on this dataset using 5-fold cross validation. In each fold, videos from four individuals are considered for training and the fifth for testing. Sample frames are shown in Fig. 4.

We compare the proposed LE-BoW approach against 3 state-of-the-art human action classification systems: (i) Laptev et al. [20], (ii) Matikainen et al. [25], (iii) Messing et al. [26]. In [25], a method for augmenting quantised local features with relative spatial-temporal relationships between pairs of features is proposed. Their discriminative classifier is trained by estimating all of the cross probabilities for various local features of an action. Messing et al. [26] track Harris3D [19] interest points with a KLT tracker [23] and extract velocity history information along the trajectories. Appearance and location features are utilised in a generative mixture model to improve the recognition performance.

Table 3 shows that the proposed LE-BoW approach with simple HA encoding outperforms the state-of-the-art methods. SC encoding obtains the highest performance in which the correct recognition accuracies for individual subjects are: 86.7, 90.0, 93.3, 90.0, 96.7. In total, SC encoding was unable to correctly classify only 13 videos (out of 150). STP encoding, with the mean correct classification rate of 90.7%, is on par with SC encoding.

Table 3: Comparisons between the proposed approach to the state-of-the-art methods on Activity of Daily Living dataset; CCR: Correct Classification Rate (in %).

| Method | CCR |
|------------------------------|-------------|
| Laptev et al. [20] | 80 |
| Messing et al. [26] | 89 |
| Matikainen et al. [25] | 70 |
| proposed LE-BoW (HA) | 90.0 |
| proposed LE-BoW (STP) | 90.7 |
| proposed LE-BoW (SC) | 91.3 |

5.4 Trade-Off Between Accuracy and Complexity

Encoding local descriptors via HA (Section 4.2.1) has the least computational burden. However, it loses the structure information between the set of covariance matrices extracted from a query video. Notable improvements are observed by encoding structure information using STP (Section 4.2.2) or SC (Section 4.2.3). Compared to both HA and STP, obtaining the histogram using SC is considerably more computationally demanding, as SC requires solving a minimisation problem for each spatio-temporal block.

The experiments show that STP provides a good trade-off between the recognition accuracy and computational cost. For example, by averaging over 10 runs and 100 videos, histogram generation using SC for one video took 33 seconds on a 2.5GHz Intel i5 CPU with 4GB of RAM using Matlab, while the same test just needed 1.2 and 0.6 seconds for STP and HA, respectively.

5.5 Further Discussion

In this subsection we provide further insight into the performance and properties of the proposed method. We first provide more performance results using Recall and Precision on the studied datasets. We then assess the effect of various spatio-temporal grids in the HOF descriptor on recognition accuracy.

Tables 4 to 6 show the performance of the proposed LE-BoW using STP encoding in terms of Precision and Recall measures. On the KTH dataset, the “boxing” action obtained the highest performance according to both of the measures. The “hand clapping” action obtained the second best Recall value, while “hand waving” obtained the second best Precision value. On the Olympic Sports dataset, none of the actions were confused with “javelin throw”. On the Activity of Daily Living dataset, “drink water” and “lookup in phonebook” were reliably recognised, i.e., 100% on both measures.

We study the effect of several spatio-temporal grids in HOF computation on recognition accuracy of KTH actions. Fig. 5 shows the performance for various values of $n_\sigma \times n_\sigma \times n_\tau$. Starting from the grid $1 \times 1 \times 2$ to $2 \times 2 \times 2$, the classification accuracy improves with further increasing the number of cells. However, there is a notable drop by moving from the grid $2 \times 2 \times 2$ to $2 \times 2 \times 3$. This is not surprising because we reject blocks with the trajectory numbers less than the dimension of HOF (i.e. 108 in this structure, $2 \times 2 \times 3 \times 9 = 108$). We observed that many blocks are rejected with the threshold value equal to 108. As a result, the final histogram is not rich enough (compared to the grid $2 \times 2 \times 2$). Further increasing in the number of spatial cells, i.e. $n_\sigma = 3$ does not yield better results and hence we opted for the grid $2 \times 2 \times 2$ in all our experiments.

Table 4: Precision and Recall on KTH dataset with STP encoding (values are in %).

| | Boxing | Hand Clapping | Hand Waving | Jogging | Running | Walking |
|------------------|--------|---------------|-------------|---------|---------|---------|
| Precision | 98.6 | 94.6 | 97.1 | 95.1 | 93.8 | 95.2 |
| Recall | 97.9 | 97.2 | 95.1 | 93.8 | 94.4 | 95.8 |

Table 5: Precision and Recall on Olympic Sports dataset with STP encoding (values are in %).

| | Basketball Layup | Bowling | Clean Jerk | Discus Throw | Hammer Throw | High Jump |
|------------------|------------------|---------|------------|--------------|--------------|-----------|
| Precision | 80.0 | 80.0 | 77.8 | 76.9 | 75.0 | 83.3 |
| Recall | 80.0 | 88.9 | 70.0 | 90.9 | 75.0 | 90.9 |

| | Javelin Throw | Long Jump | Platform 10m | Pole Vault | Shot Put | Snatch | Springboard 3m |
|-----------|---------------|-----------|--------------|------------|----------|--------|----------------|
| Precision | 100 | 71.4 | 77.8 | 85.7 | 88.9 | 77.8 | 85.7 |
| Recall | 75.0 | 83.3 | 77.8 | 75.0 | 80.0 | 77.8 | 75.0 |

| | Tennis Serve | Triple Jump | Vault |
|------------------|--------------|-------------|-------|
| Precision | 85.7 | 75.0 | 90.0 |
| Recall | 85.7 | 75.0 | 90.0 |

Table 6: Precision and Recall on Activity of Daily Living dataset with STP encoding (values are in %).

| | Answer Phone | Chop Banana | Dial Phone | Drink Water | Eat Banana | Eat Snack |
|------------------|--------------|-------------|------------|-------------|------------|-----------|
| Precision | 86.7 | 86.7 | 86.7 | 100 | 81.3 | 86.7 |
| Recall | 86.7 | 86.7 | 86.7 | 100 | 86.7 | 86.7 |

| | Lookup In Phonebook | Peel Banana | Use Silverware | Write on Whiteboard |
|------------------|---------------------|-------------|----------------|---------------------|
| Precision | 100 | 92.9 | 93.3 | 93.3 |
| Recall | 100 | 86.7 | 93.3 | 93.3 |

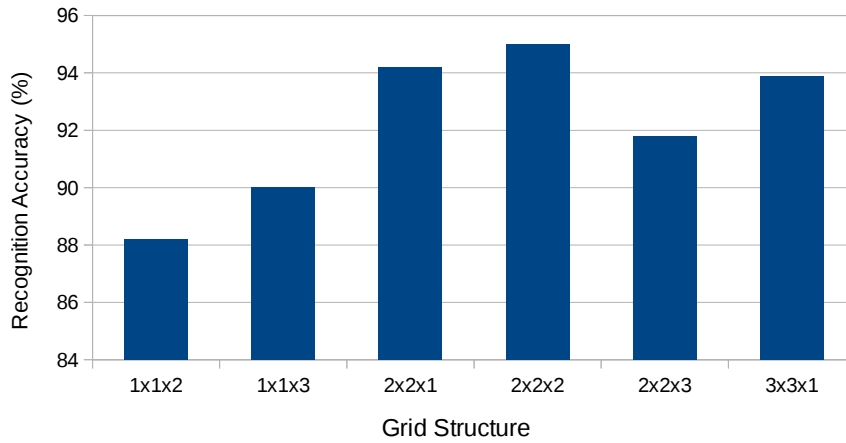


Figure 5: Evaluation of cell grid structure parameters on KTH dataset.

6 Conclusions

We devised an approach to extend the popular Bag of Words (BoW) models to a special class of non-Euclidean spaces, the space of Symmetric Positive Definite (SPD) matrices formed by covariance descriptors of spatio-temporal features [36]. In doing so, we elaborated on how a codebook and subsequently histograms can be obtained for covariance matrices and devised Log-Euclidean Bag of Words (LE-BoW), an extrinsic extension of conventional BoW using Riemannian geometry of SPD matrices. The main ingredient of our proposal is a diffeomorphism that embeds Riemannian manifold of SPD matrices into an Euclidean space. This is consistent with several studies [4, 7, 15, 12] that demonstrate the benefit of such embedding.

The proposed framework was validated by experiments on three challenging action recognition datasets, namely KTH [37], Olympic Sports [27], and Activity of Daily Living [26]. The experiments show that the proposed LE-BoW approach for classifying human actions performs better than the state-of-the-art methods proposed by Laptev et al. [20], Niebles et al. [27], and Wang et al. [44]. We believe that our work motivates future research on extending well-known machine learning inference tools to their Riemannian counterparts.

Acknowledgements

NICTA is funded by the Australian Government through the Department of Communications, as well as the Australian Research Council through the ICT Centre of Excellence program.

References

- [1] J. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3):1–16, 2011.
- [2] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.
- [3] S. Ali and M. Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):288–303, 2010.
- [4] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347, 2007.
- [5] R. Bhatia. *Positive Definite Matrices*. Princeton University Press, 2007.
- [6] C. Bishop. *Pattern recognition and machine learning*. Springer New York, 2006.
- [7] R. Caseiro, P. Martins, J. F. Henriques, and J. Batista. A nonparametric Riemannian framework on tensor field with application to foreground segmentation. *Pattern Recognition*, 45(11):3997–4017, 2012.
- [8] A. Cherian, S. Sra, A. Banerjee, and N. Papanikolopoulos. Jensen-bregman logdet divergence with application to efficient similarity search for covariance matrices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2161–2174, 2012.
- [9] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proc. Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, 2005.

- [10] M. Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer, 2010.
- [11] A. Gilbert, J. Illingworth, and R. Bowden. Action recognition using mined hierarchical compound features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):883–897, 2011.
- [12] K. Guo, P. Ishwar, and J. Konrad. Action recognition from video using feature covariance matrices. *IEEE Transactions on Image Processing*, 22(6):2479–2494, 2013.
- [13] M. T. Harandi, C. Sanderson, R. Hartley, and B. C. Lovell. Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach. In *Proc. European Conference on Computer Vision*, pages 216–229. Springer, 2012.
- [14] M. T. Harandi, C. Sanderson, A. Wiliem, and B. C. Lovell. Kernel analysis over Riemannian manifolds for visual recognition of actions, pedestrians and textures. In *Proc. IEEE Workshop on the Applications of Computer Vision*, pages 433–439, 2012.
- [15] W. Hu, X. Li, W. Luo, X. Zhang, S. Maybank, and Z. Zhang. Single and multiple object tracking using log-Euclidean Riemannian subspace and block-division appearance model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2420–2440, 2012.
- [16] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *Proc. British Machine Vision Conference*, pages 995–1004, 2008.
- [17] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2):349–396, 2003.
- [18] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.
- [19] I. Laptev and T. Lindeberg. Space-time interest points. In *Proc. Int. Conference on Computer Vision*, pages 432–439, 2003.
- [20] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.
- [21] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178, 2006.
- [22] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3337–3344, 2011.
- [23] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. Int. Joint Conference on Artificial Intelligence*, pages 674–679, 1981.
- [24] Y. M. Lui. Advances in matrix manifolds for computer vision. *Image and Vision Computing*, 30(6):380–388, 2012.
- [25] P. Matikainen, M. Hebert, and R. Sukthankar. Representing pairwise spatial and temporal relations for action recognition. In *Proc. European Conference on Computer Vision*, pages 508–521. Springer, 2010.
- [26] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *Proc. Int. Conference on Computer Vision*, pages 104–111. IEEE, 2009.
- [27] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *Proc. European Conference on Computer Vision*, pages 392–405. Springer, 2010.

- [28] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *European Conference on Computer Vision (ECCV)*, pages 490–503, 2006.
- [29] Y. Pang, Y. Yuan, and X. Li. Gabor-based region covariance matrices for face recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(7):989–993, 2008.
- [30] X. Pennec. Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006.
- [31] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010.
- [32] F. Porikli, O. Tuzel, and P. Meer. Covariance tracking using model update based on Lie algebra. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 728–735, 2006.
- [33] P. Razzaghi, M. Palhang, and N. Gheissari. A new invariant descriptor for action recognition based on spherical harmonics. *Pattern Analysis and Applications*, 16(4):507–518, 2013.
- [34] V. Reddy, C. Sanderson, and B. C. Lovell. Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 55–61, 2011.
- [35] C. Sanderson and B. C. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *Lecture Notes in Computer Science (LNCS)*, volume 5558, pages 199–208. 2009.
- [36] A. Sanin, C. Sanderson, M. T. Harandi, and B. C. Lovell. Spatio-temporal covariance descriptors for action and gesture recognition. In *Proc. IEEE Workshop on the Applications of Computer Vision*, pages 103–110, 2013.
- [37] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *Proc. Int. Conference on Pattern Recognition*, volume 3, pages 32–36, 2004.
- [38] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional SIFT descriptor and its application to action recognition. In *Proc. Int. Conference on Multimedia*, pages 357–360. ACM, 2007.
- [39] S. Sra and A. Cherian. Generalized dictionary learning for symmetric positive definite matrices with application to nearest neighbor retrieval. In *Proc. European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 318–332. Springer, 2011.
- [40] R. Subbarao and P. Meer. Nonlinear mean shift over Riemannian manifolds. *Int. Journal of Computer Vision*, 84(1):1–20, 2009.
- [41] A. Tsitsoulis and N. Bourbakis. A first stage comparative survey on human activity recognition methodologies. *International Journal on Artificial Intelligence Tools*, 22(6), 2013.
- [42] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *Proc. European Conference on Computer Vision*, pages 589–600, 2006.
- [43] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on Riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1713–1727, 2008.
- [44] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, 2013.
- [45] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference*, 2009.

- [46] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224–241, 2011.
- [47] A. Wiliem, C. Sanderson, Y. Wong, P. Hobson, R. F. Minchin, and B. C. Lovell. Automatic classification of human epithelial type 2 cell indirect immunofluorescence images using cell pyramid matching. *Pattern Recognition*, 47(7):2315–2324, 2014.
- [48] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proc. European Conference on Computer Vision*, pages 650–663. Springer, 2008.
- [49] Y. Wong, M. Harandi, and C. Sanderson. On robust face recognition via sparse coding: The good, the bad and the ugly. *IET Biometrics*, (in press). <http://dx.doi.org/10.1049/iet-bmt.2013.0033>
- [50] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.
- [51] C. Yuan, W. Hu, X. Li, S. Maybank, and G. Luo. Human action recognition under log-Euclidean Riemannian metric. In *Proc. Asian Conference on Computer Vision*, pages 343–353. Springer, 2010.